Sequenceserver: a modern graphical user interface for custom BLAST databases

Anurag Priyam [1]*, Ben J. Woodcroft [2], Vivek Rai [3], Ismail Moghul [1], Alekhya Munagala [4], Filip Ter [1], Hiten Chowdhary [4], Iwo Lukasz Pieniak [1], Mark Anthony Gibbins [5], HongKee Moon [6], Austin Davis-Richardson,[7], Mahmut Uludag [8], Nathan S. Watson-Haigh [9], Richard Challis [10], Hiroyuki Nakamura [11], Emeline Favreau [1], Esteban Gómez Cifuentes [1], Tomáš Pluskal [12], Guy Leonard [13], Wolfgang Rumpf [14], Yannick Wurm [1,15]*

[1]School of Biological and Chemical Sciences, Queen Mary University of London, London, E1 4NS, United Kingdom
[2]School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, Queensland, 4072, Australia
[3]Department of Biotechnology, Indian Institute of Technology Kharagpur, Kharagpur, 721302, India
[4]Department of Mathematics, Indian Institute of Technology Kharagpur, Kharagpur, 721302, India
[5]Department of Computer Science, Royal Holloway University of London, Surrey, TW20 0EX, United Kingdom
[6]Scientific Computing Facility, Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, 01307, Germany
[7]San Francisco, CA 94110
[8]Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia
[9]Bioinformatics Hub, School of Biological Sciences, University of Adelaide, SA5005 Adelaide, Australia
[10]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, EH9 3JT, United Kingdom
[11]Spiber Inc., 234-1 Mizukami, Kakuganji, Tsuruoka, Yamagata, 997-0052, Japan
[12]Whitehead Institute for Biomedical Research, 455 Main Street, Cambridge, MA 02142 USA
[13]Living Systems Institute, University of Exeter, Exeter, EX4 4QD, United Kingdom
[14]The Institute for Genomic Medicine, The Abigail Wexner Research Institute at Nationwide Childrens Hospital Columbus, OH. 43205, United States
[15]5Bases Limited, London, E9 7DR, United Kingdom
***Corresponding author:** E-mail: anurag.priyam@qmul.ac.uk, y.wurm@qmul.ac.uk.

Abstract

Comparing newly obtained and previously known nucleotide and amino-acid sequences underpins modern biological research. BLAST is a well-established tool for such comparisons but is challenging to use on new datasets. We combined a user-centric design philosophy with sustainable software development approaches to create Sequenceserver, a tool for running BLAST and visually inspecting BLAST results for biological interpretation. Sequenceserver uses simple algorithms to prevent potential analysis errors, and provides flexible text-based and visual outputs to support researcher productivity. Our software can be rapidly installed for use by individuals or on shared servers.

Sequenceserver is AGPLv3-licensed at https://sequenceserver.com.

Key words: visualization, BLAST, comparative genomics, sequence analysis

Brief Comm.

Introduction

The dramatic drop in sequencing costs has created many opportunities for individuals and groups of researchers to generate genomic or transcriptomic sequences from previously understudied organisms. Many research questions require small- or large-scale sequence comparisons, and BLAST (Basic Local Alignment Search Tool) is the most established tool for many such analyses (Altschul et al., 1990; Camacho et al., 2009). Unfortunately, BLAST analysis of new data can be challenging. There are delays before new data are submitted to and become publicly available on central BLAST repositories such as the NCBI (National Center for Biotechnology Information), and only small queries are feasible on such repositories. BLAST can be downloaded and installed locally, but its usage can be challenging for researchers without experience of command-line interfaces. Finally, commercial software to overcome such hurdles is too costly for many laboratories.

Here we present Sequenceserver, a free graphical interface for BLAST designed to increase the productivity of biologist researchers performing and interpreting BLAST searches on custom datasets, and of bioinformaticians setting up shared laboratory or community databases. It has a user-centric focus (Garrett, 2011) on accompanying researchers through their work process. Below, we provide an overview of Sequenceserver features that facilitate BLAST query submission and interpretation.

Assisted installation and BLAST query submission

Installing Sequenceserver on computers running macOS or Linux is typically rapid, requiring only one or few commands (see online documentation). If necessary, Sequenceserver automates the download of BLAST (Camacho et al., 2009) binaries and can manage the conversion of FASTA files to BLAST databases. A user accesses Sequenceserver's graphical interface in a web browser at http://localhost:4567 (figure 1A). All detected BLAST databases are automatically listed here. The user types, pastes or drag-and-drops FASTA format query sequences into a text-field (figure 1A). To prevent common errors, an alert message is shown and query submission is disabled if the query is invalid (e.g., combining nucleotide and protein sequences). The user then selects databases. The appropriate basic BLAST algorithm will automatically be used (figure S1, Supplementary Material online). When multiple algorithms are appropriate, a pull-down in the BLAST submission button allows the user to toggle between them. An "advanced parameters" field provides access to all standard BLAST parameters.

BLAST result visualization and further analysis

The Sequenceserver results page is designed to facilitate navigation, interpretation and follow-up analysis (figure 1B and http://sequenceserver.com/paper/resultsinteractive). Results are visually structured and

will feel familiar to users of NCBI BLAST. If multiple query sequences were submitted, a clickable index of queries is shown. Queries, hits and BLAST HSPs (high-scoring segment pairs) are numbered to facilitate navigation. For each query, identified hits are summarized in a table and an overview graphic. Each hit includes links for FASTA download, sequence visualization, and potentially to other resources. Such links can be automatically added based on regular expression analysis of identifiers (see online documentation). BLAST results can be downloaded in XML or tab-delimited table formats for further analysis. Similarly, a FASTA file containing all hit sequences, or a selection of hit sequences can be downloaded.

Usage by individual researchers and as part of community databases

Usage statistics including downloads, preprint citations, GitHub , and mailing list participation (figure 1C) indicate that Sequenceserver is extensively used for molecular-genetic research on emerging model organisms (table S1, Supplementary Material online). For example, Sequenceserver installations on personal computers helped characterize the evolution of tunicate genomes (Blanchoud et al., 2018), fire ant olfactory genes (Pracana et al., 2017) and loci affecting Sorghum shoot architecture (McCormick et al., 2016). Sequenceserver has also been used to analyze human prostate cancer genomes (Seim et al., 2017) and to identify bacteria affecting shelf life of milk (Reichler et al., 2018).

Importantly, Sequenceserver also represents a main querying mechanism for more than 50 community genome databases (table S2, Supplementary Material online), including the PHI-base database of genes underpinning pathogen-host interactions, (Winnenburg et al., 2006), an initiative to sequence 1,000 wild yeast genomes (Shen et al., 2016) and the http://reefgenomics.org coral genomics database (Liew et al., 2016). Such community resources typically integrate Sequenceserver as part of larger web servers (e.g., Nginx (Reese, 2008)) and customize it by adding links from BLAST hits to genome browsers or other gene-specific information. Additionally, many password-protected Sequenceserver instances exist for unpublished data.

Outlook

In creating Sequenceserver, we aimed to respect user-centric design principles, open-source and sustainable software engineering practices (Supplementary Material online). Our software is built using Ruby and Javascript frameworks commonly used for professional software development. The resulting robust architecture and flexibility facilitate customization and integration with other tools. This has led to contributions of improvements and bug-fixes by 21 bioinformaticians unrelated to the initial project; many are now coauthors. Our community is testing the ability to import preexisting BLAST or

DIAMOND XML result files (Buchfink et al., 2015), and new manners of visualizing results (Cui et al., 2016; Wintersinger and Wasmuth, 2015). Such efforts will continue to improve the ability of researchers to analyze and interpret genomic data.

Supplementary Material

Supplementary Material including text, tables S1-S2 and figure S1, are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org). Source code is available under GNU Affero General Public License (AGPL) 3.0 at https://github.com/sequenceserver. Additional documentation is available online at http://sequenceserver.com.

Acknowledgments

We thank the many Sequenceserver users and contributors for their input. During the creation of Sequenceserver YW was funded by an ERC grant to Laurent Keller. BJW was supported by the United States Department of Energy (DE-SC0004632). While writing this manuscript, YW and AP were supported by BBSRC (BB/K004204/1) and NERC (NE/L00626X/1).

References

Altschul S et al. (1990). Basic Local Alignment Search Tool. *J Mol Biol* 215:403–410.

Blanchoud S, Rutherford K, Zondag L, Gemmell NJ, and Wilson MJ (2018). *De novo* draft assembly of the *Botrylloides leachii* genome provides further insight into tunicate evolution. *Sci Rep* 8:5518.

Buchfink B et al. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12:59–60.

Camacho C et al. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics* 10:421.

Cui Y, Chen X, Luo H, Fan Z, Luo J, He S, Yue H, Zhang P, and Chen R (2016). BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications. *Bioinformatics* 32:1740–1742.

Garrett JJ (2011). *The elements of user experience: User-centered design for the Web and beyond*. New Riders, Berkeley.

Liew YJ, Aranda M, and Voolstra CR (2016). Reefgenomics.org. a repository for marine genomics data. *Database* baw152.

McCormick RF, Truong SK, and Mullet JE (2016). 3D sorghum reconstructions from depth images identify QTL regulating shoot architecture. *Plant Physiol* 172:823–834.

Pracana R, Levantis I, Martínez-Ruiz C, Stolle E, Priyam A, and Wurm Y (2017). Fire ant social chromosomes: Differences in number, sequence and expression of odorant binding proteins. *Evolution Lett* 1:199–210.

Reese W (2008). Nginx: The high-performance web server and reverse proxy. *Linux Journal* 173:2.

Reichler S, Trmi A, Martin N, Boor K, and Wiedmann M (2018). *Pseudomonas fluorescens* group bacterial strains are responsible for repeat and sporadic postpasteurization contamination and reduced fluid milk shelf life. *J Dairy Sci* 101:7780.

Seim I, Jeffery PL, Thomas PB, Nelson CC, and Chopin LK (2017). Whole-genome sequence of the metastatic PC3 and LNCaP human prostate cancer cell lines. *G3* 7:1731–1741.

Shen XX, Zhou X, Kominek J, Kurtzman CP, Hittinger CT, and Rokas A (2016). Reconstructing the backbone of the *Saccharomycotina* yeast phylogeny using genome-scale data. *G3* 6:3927–3939.

Winnenburg R, Baldwin TK, Urban M, Rawlings C, Köhler J, and Hammond-Kosack KE (2006). PHI-base: a new database for pathogen host interactions. *Nucleic Acids Res* 34:D459–D464.

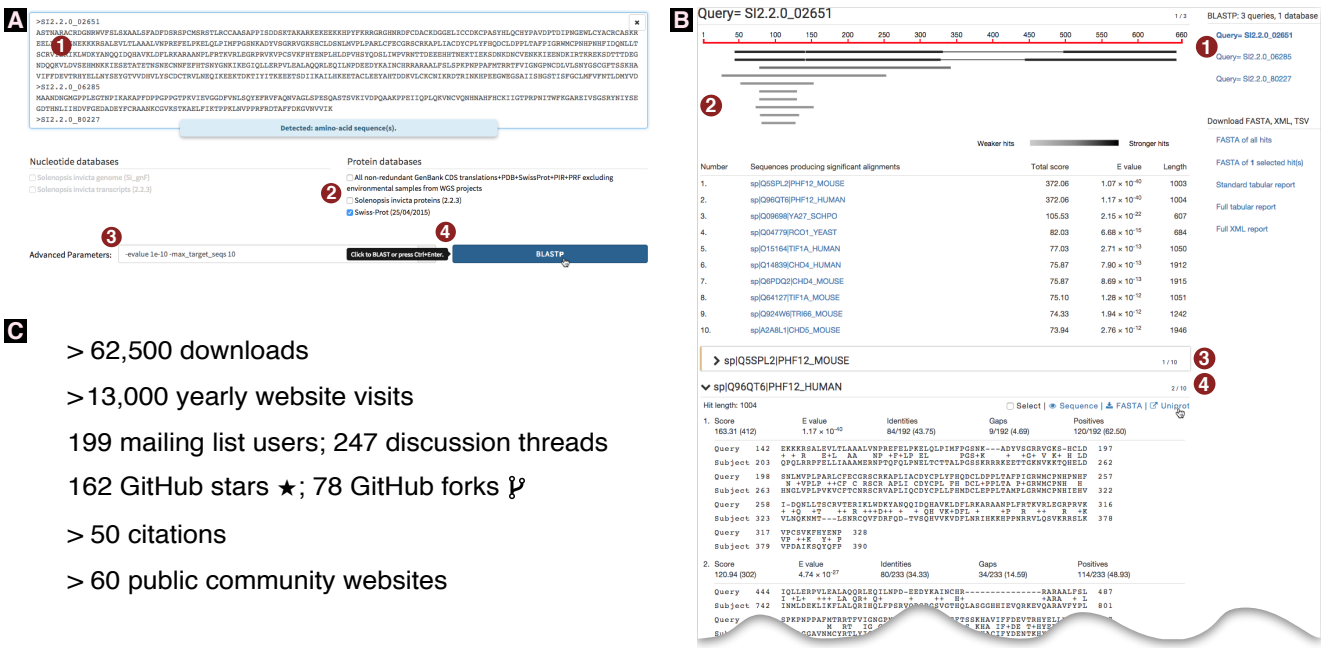

FIG. 1. A: Partial screenshot of the query interface. Numbers circled in red highlight the steps involved and some specific features. **1:** Three or more sequences were pasted into the query field (typewriter font; only the identifier is visible for the third sequence); a message confirms to the user that these are amino acid sequences. **2:** The Swiss-Prot protein database was the first database to be selected. As a result, additional database selections are limited to protein databases; nucleotide databases are disabled. **3:** Optional advanced parameters were entered which constrain the results to the 10 strongest hits with E-values stronger than 10^{-10}. **4:** The BLAST button is automatically activated and labeled "BLASTP" as this is the only possible basic BLAST algorithm for the given query-database combination. As the user's mouse pointer hovers over the BLASTP button, a tooltip indicates that a keyboard shortcut exists for this button.

B. Partial screenshot of a Sequenceserver BLAST report. An interactive version of this figure is online at http://sequenceserver.com/paper/resultsinteractive. Three amino acid sequences were compared against the Swiss-Prot database using BLASTP with an E-value cutoff of 10^{-10} and keeping only the 10 strongest hits per query. This screenshot shows a portion of the results for the first query. Numbers circled in red highlight some specific features of this report. **1:** An index overview summarizes the query and database information and provides clickable links to query-specific results. **2:** Results for the first query are shown. These include a graphical overview indicating which parts of the query sequence align to each hit, a tabular summary of all hits and alignment details for each hit. **3:** The first hit is selected for download; its alignment details have been folded away. **4:** The user is studying the second hit; the mouse pointer hovers over the link to the hit's UniProt page.

C: Sequenceserver usage as of June 11, 2019. These include download statistics from https://rubygems.org/gems/sequenceserver, Google Analytics statistics for http://sequenceserver.com, and citation statistics from https://app.dimensions.ai/details/publication/pub.1085102830, and GitHub statistics from https://github.com/wurmlab/sequenceserver.

Wintersinger JA and Wasmuth JD (2015). Kablammo: An interactive, web-based blast results visualizer. *Bioinformatics* 31:1305–1306.